# Switching nonparametric regression models for multi-curve data

**Camila P. E. de Souza,***
**Nancy E. Heckman***
**and Helena Xu***

*Department of Statistics - The University of British Columbia*
*e-mail:* camila.souza@stat.ubc.ca

**Abstract:** We develop and apply an approach for analyzing multi-curve data where each curve is driven by a latent state process. The state at any particular point determines a smooth function, forcing the individual curve to "switch" from one function to another. Thus each curve follows what we call a switching nonparametric regression model. We develop an EM algorithm to estimate the model parameters. We also obtain standard errors for the parameter estimates of the state process. We consider several types of state processes: independent and identically distributed, independent but depending on a covariate and Markov. Simulation studies show the frequentist properties of our estimates. We apply our methods to a data set of a building's power usage.

**Keywords and phrases:** EM algorithm, latent variables, machine learning, nonparametric regression, power usage, switching nonparametric regression model.

## 1. Introduction

We develop and apply a method to analyze multi-curve data where each curve follows a *switching nonparametric regression model* (De Souza and Heckman, 2014): each curve, over its domain, switches among $J$ unobserved states with each state determining a function. The main goal is to estimate the function corresponding to each state and the parameters of the latent process, along with some measure of accuracy.

We are motivated by the problem of calculating a building's "typical curve" of energy consumption, that is, its expected energy consumption as a function of time and other variables (e.g., weather conditions). Such knowledge allows building managers to compare the building's real-time performance to its "typical" performance which is useful, for instance, for assessing the impact of improvements on a building's energy efficiency. The data set we analyze was supplied to us by PulseEnergy, now part of EnerNOC.

To understand our methodological approach, compare the plots in Figures 1 and 2. Figure 1 shows hourly power usage during the months of June and July 2009 in an office building. On some days (holidays and weekends) energy usage is very close to zero. We can observe that on some business days the

*Research supported by the National Science and Engineering Research Council of Canada.

energy usage is very high, approximately twice as much as on the other days. This high power consumption occurs on warm days, when the cooling system (also called the chiller) of the building was probably on. Figure 2 presents the building daytime power usage from 9am to 4pm on 44 business days in June and July 2009. Several types of curves can be observed: one type corresponds to days when the cooling system was probably on and another type when the cooling system was off. We can also observe that on some days the chiller turned on in the middle of the day. On one day the chiller went on, off and then on again.

Brown, Barrington-Leigh and Brown (2012) consider the data in Figure 1 using a very computer intensive method. They find the "typical curve" by applying a local constant kernel smoother over an extremely large number of data points, and thus, their contribution to the analysis is mainly one of computational efficiency. They do not consider the special structure we see in Figure 2. One shortfall of their smoothing method is that they do not model the abrupt changes in level of energy consumption, and thus their approach may oversmooth these changes. Since these changes are real features in the data, they should be modelled explicitly to better understand power usage. Our method exploits the structure of Figure 2 and differs from the approach proposed by Brown, Barrington-Leigh and Brown in two important ways: one, by treating each business day as a replicate and the other, by modelling abrupt changes by modelling the building power usage as arising from two functions, one function giving power usage when the chiller is off, the other function giving power usage when the chiller is on. The condition "chiller on"/"off" at any particular time cannot be observed directly. It can only be inferred from the data, thus forming a latent process.

De Souza and Heckman (2014) present the case where there is a single realization, a single curve switching among $J$ functions. In that paper, we consider two models for the latent process: one where the states are independent and identically distributed, the other where the sequence of states forms a Markov chain. In addition to estimating all parameters and functions, we derive standard errors for the parameters of the latent process. In the present paper we extend our 2014 approach, considering the case when there are $N$ curves, called replicates, with each replicate switching among $J$ functions. We also consider a third type of latent state process, where the state depends on a time-varying covariate.

Several authors have considered the single realization case, but from a Bayesian perspective with the smooth functions modeled as realizations of Gaussian processes. See, for instance, Tresp (2001), Rasmussen and Ghahramani (2002) and Ou and Martin (2008). These papers are discussed in more detail in De Souza and Heckman (2014). The paper of Ou and Martin (2008) also contains a Bayesian analysis of the replicate case. These three papers contain methodology that can, in principle, lead to estimation of all functions and the latent variable process parameters. However, unlike our work, these three papers focus on estimation of just one function - the mixture.

In a more recent related work, Langrock et al. (2015) consider generalized

additive models with a time component, where the predictor is subject to regime changes controlled by an underlying Markov process. The parameter estimates are obtained by a numerical maximum penalized likelihood approach. These authors focus on a single realization case and do not consider the replicate case.

## 2. Overview of the proposed methodology

We consider a data set with $N$ replicates where replicate $k$ contains $n$ observations $y_{1k}, \ldots, y_{nk}$ and evaluation points $x_1, \ldots, x_n$, which for simplicity are the same across replicates. Observation $y_{ik}$ depends on $x_i$ according to a hidden (unobserved) state $z_{ik}$ with possible state values in $\{1, \ldots, J\}$. If $z_{ik} = j$ the expected response of $y_{ik}$ is $f_j(x_i)$. In this work, we assume the replicates are all generated from just one set of functions $f_1, \ldots, f_J$, a reasonable assumption for the power usage data presented in Figure 2 and described in Section 1. We consider three types of hidden states, those that are independent and identically distributed, those that follow a Markov structure and those that are independent but with distribution depending on some covariate(s).

Our notation is as follows (with replicate index being $k = 1, \ldots, N$).

- Observed data: $\mathbf{x} = (x_1, \ldots, x_n)^T$, fixed across replicates; covariate vectors $\mathbf{v}_{1k}, \ldots, \mathbf{v}_{nk}$; responses $\mathbf{y}^R = (\mathbf{y}_1^T, \ldots, \mathbf{y}_N^T)^T$, where $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})^T$.
- Hidden states: $\mathbf{z}^R = (\mathbf{z}_1^T, \ldots, \mathbf{z}_N^T)^T$, where $\mathbf{z}_k = (z_{1k}, \ldots, z_{nk})^T$, and $z_{ik}$ takes a value in $\{1, \ldots, J\}$.
- $f_j(\mathbf{x}) = (f_j(x_1), \ldots, f_j(x_n))^T$ for $j = 1, \ldots, J$, and $f_{\mathbf{z}_k}(\mathbf{x}) = \big(f_{z_{1k}}(x_1), \ldots, f_{z_{nk}}(x_n)\big)^T$.

We assume that the replicates, $\mathbf{z}_1, \ldots, \mathbf{z}_N$, are independent. Given the hidden states $\mathbf{z}_k$, $\mathbf{y}_k = f_{\mathbf{z}_k}(\mathbf{x}) + \boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N$, are independent and $\boldsymbol{\epsilon}_k$ has a multivariate normal distribution with mean equal to the 0-vector and covariance matrix $\mathbf{V}$, possibly depending on $\mathbf{z}_k$. We denote this $\boldsymbol{\epsilon}_k \sim MVN(\mathbf{0}, \mathbf{V})$. Therefore, $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent and, given the hidden states $\mathbf{z}_k$, $\mathbf{y}_k \sim MVN(f_{\mathbf{z}_k}(\mathbf{x}), \mathbf{V})$. We let $\gamma$ be the set containing $f_1(\mathbf{x}), \ldots, f_J(\mathbf{x})$ and the parameters in $\mathbf{V}$. We assume that the distribution of each $\mathbf{z}_k$ is governed by a parameter vector $\alpha$. Section 2.1 presents our different choices of $\mathbf{V}$ and $\alpha$.

Our goal is to estimate $\theta \equiv \{\alpha, \gamma\}$, along with standard errors or some measure of accuracy for the parameters in $\alpha$. Similar to De Souza and Heckman (2014) we obtain the parameter estimates by maximizing

$$l(\theta) \equiv \sum_{k=1}^{N} \log p(\mathbf{y}_k|\theta) + P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J), \tag{1}$$

where $p(\mathbf{y}_k|\theta)$ is the likelihood function based on the observed data from the $k$th replicate and the exact form of $P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J)$ is chosen by the user. For our work, we set

$$P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J) = -\sum_{j=1}^{J} \lambda_j \int [f_j''(x)]^2 dx. \tag{2}$$

The $\lambda_j$'s are the smoothing parameters.

The form of $\log p(\mathbf{y}_k|\theta)$ is very complicated, since it involves the distribution of the latent $z_{ik}$'s. Therefore, we apply an Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) to maximize (1). We can show (see, for instance, Cappé, Moulines and Rydén, 2005 and McLachlan and Krishnan, 2008) that our EM algorithm generates a sequence of estimates, $\theta^{(c)}$, $c \geq 1$, satisfying $l(\theta^{(c+1)}) \geq l(\theta^{(c)})$.

As in De Souza and Heckman (2014) one could also take a Bayesian approach by maximizing (1) with $P$ arising from placing a Gaussian process prior on the $f_j$'s.

### 2.1. Choices of $\mathbf{V}$ and $\boldsymbol{\alpha}$

We consider five models for $\mathbf{V}$, the covariance of the residual error: $\mathbf{V}$ unrestricted, $\mathbf{V}$ diagonal with either $\mathbf{V} = \sigma^2\mathbf{I}$ or with $ii$th entry depending on the latent state, and two models generated from a "random intercept" covariance structure: a *homogeneous random intercept model* and a *non-homogeneous random intercept model*, the latter with variability of the intercept depending on the value of the latent state. We will usually write $\mathbf{V_z}$ in models where the variability depends on the latent state. However, sometimes we will omit the subscript $\mathbf{z}$ when referring to a general $\mathbf{V}$. The unknown parameters in $\mathbf{V}$ are clear for our first two models. For the third model, the parameters in $\mathbf{V_z}$ are $\sigma_1^2, \ldots, \sigma_J^2$.

To define $\mathbf{V}$ following a homogeneous random intercept model, let $y_{ik} = f_{z_{ik}}(x_i) + \epsilon_{ik}$. Suppose that

- $\epsilon_{ik} = \delta_k + e_{ik}$;
- $\{\delta_1, \ldots, \delta_N\}$ and $\{e_{ik}$ for $i = 1, \ldots, n,\ k = 1, \ldots, N\}$ are independent;
- $\delta_k$'s are independent and identically distributed ($iid$) $N(0, \tau^2)$;
- $e_{ik}$'s are $iid$ $N(0, \sigma^2)$.

Therefore, $\mathbf{V}$ depends on only two parameters and can be written as

$$\mathbf{V} = \sigma^2\big(\mathbf{I} + d\mathbf{1}\mathbf{1}^T\big), \tag{3}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix, $\mathbf{1}$ is an $n$-vector of ones and $d = \tau^2/\sigma^2$.

Our data analysis (Section 6) requires the more complex covariance structure of a non-homogenous random intercept model, where the variance of the random intercept depends on the state. We define this model for the simple case, where there are $J = 2$ states. We assume that $y_{ik} = f_{z_{ik}}(x_i) + \epsilon_{z_{ik},\,ik}$, where

- if $z_{ik} = 1$ then $\epsilon_{1,ik} = \delta_k + e_{ik}$;
- if $z_{ik} = 2$ then $\epsilon_{2,ik} = \delta_k + \vartheta_k + e_{ik}$;
- $\delta_k$, $\vartheta_k$, and $e_{ik}$ are independent for $i = 1, \ldots, n$ and $k = 1, \ldots, N$;
- $\delta_k$'s are $iid$ $N(0, \tau_1^2)$;
- $\vartheta_k$'s are $iid$ $N(0, \tau_2^2)$;
- $e_{ik}$'s are $iid$ $N(0, \sigma^2)$.

Therefore, the covariance matrix for the non-homogeneous random intercept model is given by

$$\mathbf{V}_{\mathbf{z}_k} = \sigma^2(\mathbf{I} + d_1\mathbf{1}\mathbf{1}^T + d_2\mathbf{1}_{\mathbf{z}_k}\mathbf{1}_{\mathbf{z}_k}^T), \tag{4}$$

where $d_j = \tau_j^2/\sigma^2$ and $\mathbf{1}_{\mathbf{z}_k}$ is an $n$-vector with $i$th entry $\mathrm{I}(z_{ik} = 2)$.

In our model $\alpha$ is the vector containing the parameters governing the distribution of the hidden states. If the $z_{ik}$'s are *iid*, then $\alpha$ is of length $J$ with $j$th component equal to $p(z_{ik} = j) \equiv p_j$. If the $z_{ik}$'s follow a Markov structure, that is, if $p(z_{ik}|z_{(i-1)k}, \ldots, z_{1k}, \alpha) = p(z_{ik}|z_{(i-1)k}, \alpha)$, $i = 2, \ldots N$, then the parameter vector $\alpha$ consists of the initial probabilities, $\pi_j = p(z_{ik} = 1|\alpha)$, and the transition probabilities, $a_{lj} = p(z_{ik} = j|z_{(i-1)k} = l, \alpha)$, $j, l = 1, \ldots, J$. Note that the transition probabilities do not depend on $i$ or $k$.

In the case where the $z_{ik}$'s are independent, with $z_{ik}$'s distribution depending on a vector of covariates $\mathbf{v}_{ik} = (1, v_{1,ik}, v_{2,ik}, \ldots, v_{M,ik})^T$, we assume that $p(z_{ik} = j|\mathbf{v}_{ik}) \equiv p_j(\mathbf{v}_{ik})$ follows a multinomial logistic regression model with

$$\log\frac{p_j(\mathbf{v}_{ik})}{p_1(\mathbf{v}_{ik})} = \beta_{j0} + \beta_{j1}v_{1,ik} + \cdots + \beta_{jM}v_{M,ik} = \boldsymbol{\beta}_j^T\mathbf{v}_{ik} \text{ for } j = 2, \ldots, J$$

so that

$$p_1(\mathbf{v}_{ik}) = \frac{1}{1 + \sum_{j=2}^J e^{\boldsymbol{\beta}_j^T\mathbf{v}_{ik}}}$$

and

$$p_j(\mathbf{v}_{ik}) = \frac{e^{\boldsymbol{\beta}_j^T\mathbf{v}_{ik}}}{1 + \sum_{j=2}^J e^{\boldsymbol{\beta}_j^T\mathbf{v}_{ik}}} \text{ for } j = 2, \ldots, J.$$

In this case $\alpha$ contains all the regression coefficient vectors $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J$.

## 3. Parameter estimation

We present the EM algorithm that we use to obtain the estimates of the parameters in $\theta$. In the M-step, we take the same approach as De Souza and Heckman (2014) and model each $f_j$ as a linear combination of $K$ known cubic B-spline basis functions, so that $f_j(\mathbf{x}) = \mathbf{B}\phi_j$, where $\phi_j$ is the $K$-vector of coefficients corresponding to $f_j$ and $\mathbf{B}$ is the $n \times K$ matrix with entries $B_{i\nu} = b_\nu(x_i)$.

The smoothing parameters, $\lambda_1, \ldots, \lambda_J$, can be chosen by a data driven method or by eye. In Section 3.4, we propose and justify a leave-one-curve out cross-validation (CV) criterion to find the optimal $\lambda_j$'s for the case when $\mathbf{V}$ is diagonal and use this method in our application. However, in our application when we assume a non-diagonal $\mathbf{V}$ and in our simulations, we choose the $\lambda_j$'s by eye.

Let $p(\mathbf{y}^R, \mathbf{z}^R|\theta)$ be the joint distribution of the observed and latent data given $\theta$, also called the complete data distribution. The application of the EM algorithm to the replicate case is similar to that in the one realization case in De Souza and Heckman (2014). It is based on writing

$$\log p(\mathbf{y}^R, \mathbf{z}^R|\theta) = \log p(\mathbf{y}^R|\mathbf{z}^R, \theta) + \log p(\mathbf{z}^R|\theta) \equiv \mathcal{L}_1(\gamma) + \mathcal{L}_2(\alpha).$$

In the E-step we calculate

$$Q(\theta, \theta^{(c)}) \equiv \mathrm{E}_{\theta^{(c)}}\left(\log p(\mathbf{y}^R, \mathbf{z}^R|\theta)|\mathbf{y}^R\right) = \mathrm{E}_{\theta^{(c)}}(\mathcal{L}_1(\gamma)|\mathbf{y}^R) + \mathrm{E}_{\theta^{(c)}}(\mathcal{L}_2(\alpha)|\mathbf{y}^R). \tag{5}$$

In the M-step, we maximize $S(\theta, \theta^{(c)}) \equiv Q(\theta, \theta^{(c)}) + P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J)$ as a function of $\theta$, to find our updated $\theta^{(c+1)}$.

### *3.1. E-step*

Let $\mathbf{s}$ be an $n$-vector of possible hidden states, i.e., each entry of $\mathbf{s}$ is in $\{1, 2, \ldots, J\}$, and let

$$p_k(\mathbf{s})^{(c)} \equiv p(\mathbf{z}_k = \mathbf{s}|\mathbf{y}^R, \theta^{(c)}) = p(\mathbf{z}_k = \mathbf{s}|\mathbf{y}_k, \theta^{(c)}).$$

Then, since

$$\mathcal{L}_1(\gamma) = \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} \mathrm{I}(\mathbf{z}_k = \mathbf{s}) \log p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{s}, f_{\mathbf{s}}(\mathbf{x}), \mathbf{V}_{\mathbf{s}}),$$

$$\mathrm{E}_{\theta^{(c)}}(\mathcal{L}_1(\gamma)|\mathbf{y}^R)) = \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \log p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{s}, f_{\mathbf{s}}(\mathbf{x}), \mathbf{V}_{\mathbf{s}})$$

$$= -\frac{Nn}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \left[(\mathbf{y}_k - f_{\mathbf{s}}(\mathbf{x}))^T \mathbf{V}_{\mathbf{s}}^{-1}(\mathbf{y}_k - f_{\mathbf{s}}(\mathbf{x})) + \log|\mathbf{V}_{\mathbf{s}}|\right]. \tag{6}$$

Similarly, we can write

$$E_{\theta^{(c)}}(\mathcal{L}_2(\alpha)|\mathbf{y}) = \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \log p(\mathbf{z}_k = \mathbf{s}|\alpha). \tag{7}$$

It is important to note that, in the building daytime power usage application, $n$ (the length of $\mathbf{s}$) is small so that, fortunately, in (6) and (7), the sum over all possible $\mathbf{s}$'s is not too large. Note also that the calculation of $\mathrm{E}_{\theta^{(c)}}(\mathcal{L}_1(\gamma)|\mathbf{y}^R)$ depends on the model for the $z_{ik}$'s only via $p_k(\mathbf{s})^{(c)}$, while $E_{\theta^{(c)}}(\mathcal{L}_2(\alpha)$ also depends on $\log p(\mathbf{z}_k = \mathbf{s}|\alpha)$.

Calculation of $p_k(\mathbf{s})^{(c)}$ is easy, by first writing

$$\begin{aligned} p_k(\mathbf{s})^{(c)} &= \frac{p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{s}, \theta^{(c)}) \times p(\mathbf{z}_k = \mathbf{s}|\theta^{(c)})}{p(\mathbf{y}_k|\theta^{(c)})} \\ &= \frac{p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{s}, f_{\mathbf{s}}(\mathbf{x})^{(c)}, \mathbf{V}^{(c)}) \times p(\mathbf{z}_k = \mathbf{s}|\alpha^{(c)})}{\sum_{\text{all } \mathbf{u}\text{'s}} p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{u}, f_{\mathbf{u}}(\mathbf{x})^{(c)}, \mathbf{V}^{(c)}) \times p(\mathbf{z}_k = \mathbf{u}|\alpha^{(c)})}. \end{aligned}$$

We calculate $p(\mathbf{y}_k|\mathbf{z}_k = \mathbf{s}, f_{\mathbf{s}}(\mathbf{x}), \mathbf{V}^{(c)})$ using the normality assumption for the distribution of $\mathbf{y}_k$ given the hidden states $\mathbf{z}_k$. Thus, we need only calculate $p(\mathbf{z}_k = \mathbf{s}|\alpha)$ for each latent state model, which is straightforward.

### 3.2. M-step via an ECM algorithm

From (5) and the calculations in Section 3.1, we see that we want to find $\theta^{(c+1)}$ that maximizes

$$S^*(\theta, \theta^{(c)}) \equiv$$

$$- \quad \frac{1}{2}\sum_{k=1}^{N}\sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \left[(\mathbf{y}_k - f_{\mathbf{s}}(\mathbf{x}))^T \mathbf{V}_{\mathbf{s}}^{-1}(\mathbf{y}_k - f_{\mathbf{s}}(\mathbf{x})) + \log|\mathbf{V}_{\mathbf{s}}|\right] \tag{8}$$

$$+ \quad P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J) \tag{9}$$

$$+ \quad \sum_{k=1}^{N}\sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \log p(\mathbf{z}_k = \mathbf{s}|\alpha). \tag{10}$$

with respect to $\theta = \{\alpha, f_1(\mathbf{x}), \ldots, f_J(\mathbf{x})$ and the parameters in $\mathbf{V}\}$, or at least satisfies $S^*(\theta^{(c+1)}, \theta^{(c)}) \geq S^*(\theta^{(c)}, \theta^{(c)})$. Note that $\theta^{(c)}$ is fixed and thus so are the $p_k(\mathbf{s})^{(c)}$'s. We also consider the smoothing parameters, $\lambda_1, \ldots, \lambda_J$, to be fixed. We apply a natural extension of the EM approach, the Expectation-Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), to find the sequence of $\theta^{(c)}$'s.

Because (10) does not depend on the $f_j$'s or $\mathbf{V}$, the $f_j$'s and $\mathbf{V}$ that maximize $S^*$ are the $f_j$'s and $\mathbf{V}$ that maximize (8) + (9). Therefore, the form of the maximizing $f_j$'s and $\mathbf{V}$ will only depend on the model for the $z$'s via the $p_k(\mathbf{s})^{(c)}$'s.

The steps of the ECM algorithm can be summarized as follows.

1. Hold $\mathbf{V}$ and the parameters in $\alpha$ fixed and maximize (8) + (9) with respect to the $f_j(\mathbf{x})$'s, obtaining $f_1(\mathbf{x})^{(c+1)}, \ldots, f_J(\mathbf{x})^{(c+1)}$.
2. Hold the $f_j(\mathbf{x})$'s and the parameters in $\alpha$ fixed and maximize (8) with respect to the parameters in $\mathbf{V}$, obtaining $\mathbf{V}^{(c+1)}$.
3. Now hold the $f_j(\mathbf{x})$'s and $\mathbf{V}$ fixed and maximize (10) with respect to the parameters in $\alpha$, obtaining $\alpha^{(c+1)}$.

We now present details of the calculations of steps 1 to 3.

1. *Updating the $f_j(\mathbf{x})$'s (any $\mathbf{V}$).*
   We propose a method to update the $f_j(\mathbf{x})$'s that is straightforward and yields an estimate of $\mathbf{f} = (f_1(\mathbf{x})^T, \ldots, f_J(\mathbf{x})^T)^T$ in a closed form. The trick is to write $f_{\mathbf{s}}(\mathbf{x})$ in terms of $f_1(\mathbf{x}), \ldots, f_J(\mathbf{x})$. To do this, let $\mathbf{1}_{j,\mathbf{s}}$ be the $n$-vector with $i$th element equal to 1 if $s_i = j$, 0 else. Let $\mathcal{I}_{\mathbf{s}}$ be the $n$ by $nJ$ matrix, $\mathcal{I}_{\mathbf{s}} = [\ \text{diag}(\mathbf{1}_{1,\mathbf{s}}) \ \big|\ \cdots \ \big|\ \text{diag}(\mathbf{1}_{J,\mathbf{s}})]$. Then we easily see that $f_{\mathbf{s}}(\mathbf{x}) = \mathcal{I}_{\mathbf{s}}\mathbf{f}$. Recall that $\mathbf{f}_j(\mathbf{x}) = \mathbf{B}\phi_j$. Let $\mathbf{B}^*$ be the $nJ \times KJ$ block diagonal matrix with each block equal to $\mathbf{B}$ and let $\phi$ be the $JK$-vector $\phi = (\phi_1^T, \ldots, \phi_J^T)^T$. Therefore $\mathbf{f} = \mathbf{B}^*\phi$. Let $\mathbf{R}$ be the $K \times K$ matrix with entries $\mathbf{R}_{\nu\nu'} = \int b''_\nu(x) b''_{\nu'}(x)\, dx$. Combining these calculations we see that, to find the $f_j$'s that maximize (8) + (9), we must maximize, as a function

of $\phi$,

$$-\frac{1}{2}\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)}\left[(\mathbf{y}_k-\mathcal{I}_{\mathbf{s}}\mathbf{B}^*\phi)^T\mathbf{V}_{\mathbf{s}}^{-1}(\mathbf{y}_k-\mathcal{I}_{\mathbf{s}}\mathbf{B}^*\phi)\right]-\phi^T\text{diag}(\lambda_1\mathbf{R},\ldots,\lambda_J\mathbf{R})\phi.$$

This expression is quadratic in $\phi$ and is easily maximized in closed form. Let $\phi^{(c+1)}$ be this maximizing $\phi$ when we set $\mathbf{V}=\mathbf{V}^{(c)}$. So we let $\mathbf{f}^{(c+1)}=\mathbf{B}^*\phi^{(c+1)}$.

2. *Updating* $\mathbf{V}$.
For a model with $\mathbf{V}_{\mathbf{s}}\equiv\mathbf{V}$, with no dependence on the state vector $\mathbf{s}$ and no restrictions on the form of $\mathbf{V}$, we show in Section 1 of the Supplementary Material that $\mathbf{V}^{(c+1)}$ is

$$\widehat{\mathbf{V}}=\frac{1}{N}\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)}\big(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x})\big)\big(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x})\big)^T. \tag{11}$$

with $f_{\mathbf{s}}(\mathbf{x})=f_{\mathbf{s}}(\mathbf{x})^{(c+1)}$. Note that if the values of $\mathbf{z}_k$ were non-random and known, then $p_k(\mathbf{s})^{(c)}$ is a delta function and so $\widehat{\mathbf{V}}$ is similar to the sample covariance matrix of the $\mathbf{y}_k$'s.

When $\mathbf{V}_{\mathbf{s}}\equiv\mathbf{V}$ follows a homogeneous random intercept model, according to our calculations in Section 1 of the Supplementary Material, we update the parameter estimates of the restricted $\mathbf{V}$ in (3) as follows. Let $\sigma^{2\,(c+1)}$ be

$$\begin{aligned}\hat{\sigma}^2 \;=\; &\frac{1}{N(n-1)}\left(\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)}(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x}))^T(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x}))\right.\\ &\left.-\frac{1}{n}\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)}\left[\big(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x})\big)^T\mathbf{1}\right]^2\right),\end{aligned} \tag{12}$$

and $d^{(c+1)}$ be

$$\hat{d}=\frac{1}{\sigma^2Nn^2}\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)}\left[\big(\mathbf{y}_k-f_{\mathbf{s}}(\mathbf{x})\big)^T\mathbf{1}\right]^2-\frac{1}{n} \tag{13}$$

with $\sigma^2$ replaced by $\sigma^{2\,(c+1)}$. Therefore $\tau^{2\,(c+1)}=d^{(c+1)}\times\sigma^{2\,(c+1)}$.

The maximization in step 2 when $\mathbf{V}_{\mathbf{s}}$ follows the non-homogeneous random intercept model is given in Section 1 of the Supplementary Material, for the case that $J=2$ states. The ECM algorithm for diagonal $\mathbf{V}_{\mathbf{s}}$ is given in Section 3.3.

3. *Updating* $\alpha$ *(any* $\mathbf{V}$*).*
We must maximize (10) with respect to $\alpha$, with the calculations depending on the proposed model for the hidden states.

For *iid* $z_{ik}$'s, $p_j = p(z_{ik} = j|\alpha)$ and (10) becomes

$$\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \sum_{i=1}^{n} \sum_{j=1}^{J} \log p_j \; \mathrm{I}\{s_i = j\} = \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \sum_{j=1}^{J} \; n_{\mathbf{s},j} \log p_j$$

where $n_{\mathbf{s},j} = \sum_{i=1}^{n} \mathrm{I}\{s_i = j\}$, i.e., the number of entries in $\mathbf{s}$ that are equal to $j$. We maximize this as a function of $p_1, \ldots, p_J$, using a Lagrange multiplier for the restriction that $\sum_{j=1}^{J} p_j = 1$, obtaining:

$$p_j^{(c+1)} = \frac{1}{Nn} \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} n_{\mathbf{s},j}.$$

For Markov $z_{ik}$'s, where the vector $\alpha$ is composed of transition probabilities $a_{lj}$ and initial probabilities $\pi_j$, we rewrite (10) as

$$\begin{aligned} &\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \left[ \log \pi_{s_1} + \sum_{i=2}^{n} \log a_{s_{i-1} s_i} \right] \\ &= \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \left[ \sum_{j=1}^{J} \mathrm{I}(s_1 = j) \log \pi_j + \sum_{j=1}^{J} \sum_{l=1}^{J} n_{\mathbf{s},lj} \log a_{lj} \right] \end{aligned}$$

where $n_{\mathbf{s},lj}$ is the number of transitions in $\mathbf{s}$ from state $l$ to state $j$, that is, $n_{\mathbf{s},lj} = \sum_{i=2}^{n} \mathrm{I}\{s_{i-1} = l, s_i = j\}$. We first maximize this with respect to $a_{lj}$, holding $\pi_j$ fixed, and then maximize with respect to $\pi_j$, holding the $a_{lj}$'s fixed. Using Lagrange multipliers for the constraints $\sum_{j=1}^{J} a_{lj} = 1$ and $\sum_{j=1}^{J} \pi_j = 1$, we obtain

$$a_{lj}^{(c+1)} = \frac{\displaystyle\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} n_{\mathbf{s},lj}}{\displaystyle\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \sum_{i=2}^{n} \mathrm{I}(s_{i-1} = l)}$$

$$\pi_j^{(c+1)} = \frac{1}{N} \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \mathrm{I}(s_1 = j).$$

For $z_{ik}$'s independent with distribution depending on some covariate(s), $\alpha$ contains the regression coefficients from our logistic regression model for $p(z_{ik} = j|\mathbf{v}_{ik}, \alpha) \equiv p_j(\mathbf{v}_{ik}, \alpha)$. In this case, (10) becomes

$$\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} p_k(\mathbf{s})^{(c)} \; \sum_{i=1}^{n} \sum_{j=1}^{J} \log p_j(\mathbf{v}_{ik}, \alpha) \; \mathrm{I}\{s_i = j\},$$

which must be maximized numerically, for instance, via a Newton-Raphson method.

### 3.3. ECM algorithm when V is diagonal

Recall that we consider two cases of $\mathbf{V}$ diagonal, one with $\mathbf{V} = \sigma^2\mathbf{I}$ and one with $\mathbf{V} = \mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{\mathbf{z}_{nk}})$. We could use the notation and steps of Section 3.2, modifying Step 2 for these $\mathbf{V}$'s. However, it is much easier to re-derive all three steps using the independence of the components of $\mathbf{y}_k$ to rewrite $\mathcal{L}_1(\gamma)$, and thus $S(\theta, \theta^{(c)})$, in a simpler form. We will see below that, instead of the $p_k(\mathbf{s})^{(c)}$'s in (8) and (10), we require the simpler

$$p_{ik}(j)^{(c)} = p(z_{ik} = j|\mathbf{y}_k, \theta^{(c)}).$$

The forms of $p_{ik}(j)^{(c)}$ are given in the Appendix A

Here, we carry out the calculations of the ECM algorithm for the case that $\mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{\mathbf{z}_{nk}})$, then provide the results for the case that $\mathbf{V} = \sigma^2\mathbf{I}$. For $\mathbf{V}_{\mathbf{z}_k}$, we can write

$$\begin{aligned}\mathcal{L}_1(\gamma) &= \sum_{k=1}^{N}\sum_{i=1}^{n} \log p(y_{ik}|z_{ik}; f_{z_{ik}}(x_i), \sigma^2_{z_{ik}}) \\ &= -\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{J} \mathrm{I}\{z_{ij} = j\}\left[\log(2\pi) + \log\sigma_j^2 + \frac{(y_{ik} - f_j(x_i))^2}{\sigma_j^2}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}S^*(\theta, \theta^{(c)}) &= -\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{1k}(j)^{(c)} \log\sigma_j^2 && (14)\\ &\quad -\frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{J} \big(\mathbf{y}_k - f_j(\mathbf{x})\big)^T \mathbf{W}_{kj}\big(\mathbf{y}_k - f_j(\mathbf{x})\big) && (15)\\ &\quad + \; P(f_1, \ldots, f_J, \lambda_1, \ldots, \lambda_J) && (16)\\ &\quad + \; \mathrm{E}_{\theta^{(c)}}(\mathcal{L}_2(\alpha)|\mathbf{y}^R), && (17)\end{aligned}$$

where

$$\mathbf{W}_{kj} = \sigma_j^{-2}\text{diag}(p_{1k}(j)^{(c)}, \ldots, p_{nk}(j)^{(c)}). \quad (18)$$

To maximize $S^*(\theta, \theta^{(c)})$ as a function of $\theta = \{f_j(\mathbf{x}), \sigma_j^2,\ j = 1, \ldots, J, \text{ and } \alpha\}$ we apply the ECM algorithm as follows.

1. *Updating the $f_j(\mathbf{x})$'s.* Holding the $\sigma_j^2$'s and the parameters in $\alpha$ fixed and maximizing (15) + (16) with respect to $f_j(\mathbf{x})$ we obtain

$$\hat{f}_j(\mathbf{x}) = \sum_{k=1}^{N} \mathbf{H}_{kj}(\lambda_j)\mathbf{y}_k, \quad (19)$$

where

$$\mathbf{H}_{kj}(\lambda) = \mathbf{B}\left(\mathbf{B}^T \sum_{r=1}^{N} \mathbf{W}_{rj}\,\mathbf{B} + 2\lambda\mathbf{R}\right)^{-1} \mathbf{B}^T\mathbf{W}_{kj}. \tag{20}$$

We let $f_j(\mathbf{x})^{(c+1)}$ be $\hat{f}_j(\mathbf{x})$ with $\sigma_j^2$ in $\mathbf{W}_{kj}$ replaced by $\sigma_j^{2(c)}$.

2. *Updating the $\sigma_j^2$'s.* Now holding the $f_j(\mathbf{x})$'s and $\alpha$ fixed and maximizing (14) + (15) with respect to $\sigma_j^2$ we get

$$\hat{\sigma}_j^2 = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n} p_{ik}(j)^{(c)}\left[y_{ik} - f_j(x_i)\right]^2}{\sum_{k=1}^{N}\sum_{i=1}^{n} p_{ik}(j)^{(c)}}.$$

Let $\sigma_j^{2(c+1)}$ be $\hat{\sigma}_j^2$ with $f_j(x_i) = f_j(x_i)^{(c+1)}$.

3. *Updating $\alpha$.* Now we hold the $f_j(\mathbf{x})$'s and the $\sigma_j^2$'s fixed and maximize (17) with respect to the parameters in $\alpha$. Recall that $\mathcal{L}_2(\alpha) = \log[p(\mathbf{z}^R|\alpha)]$. For *iid* $z_{ik}$'s we write

$$\begin{aligned}
\mathrm{E}_{\theta^{(c)}}(\mathcal{L}_2(\alpha)|\mathbf{y}^R) &= \mathrm{E}_{\theta^{(c)}}\left(\sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{J} \log p_j\ \mathrm{I}\{z_{ik}=j\}|\mathbf{y}^R\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{J} \log p_j\ p_{ik}(j)^{(c)}.
\end{aligned}$$

Maximizing with respect to $p_j$, with the constraint that $\sum_{j=1}^{J} p_j = 1$, yields

$$p_j^{(c+1)} = \frac{1}{Nn}\sum_{k=1}^{N}\sum_{i=1}^{n} p_{ik}(j)^{(c)}.$$

Similiarly, for Markov $z_{ik}$'s, we calculate

$$a_{lj}^{(c+1)} = \frac{\sum_{k=1}^{N}\sum_{i=2}^{n} p(z_{(i-1)k}=l, z_{ik}=j|\mathbf{y}_k, \theta^{(c)})}{\sum_{k=1}^{N}\sum_{i=2}^{n} p(z_{(i-1)k}=l|\mathbf{y}_k, \theta^{(c)})}$$

and

$$\pi_j^{(c+1)} = \frac{1}{N}\sum_{k=1}^{N} p_{1k}(j)^{(c)}.$$

For $z_{ik}$'s independent with distribution depending on some covariates we need numerical optimization methods, such as Newton-Raphson, to

obtain the coefficient estimates from our logistic regression model for $p(z_{ik} = j|\mathbf{v}_{ik}) \equiv p_j(\mathbf{v}_{ik})$. So, for example, if there are $J = 2$ states and the covariate vector is $\mathbf{v}_{ik} = (1, v_{ik})^T$, we apply a numerical method to obtain $\beta_{20}$ and $\beta_{21}$ that maximize

$$\mathrm{E}_{\theta^{(c)}}(\mathcal{L}_2(\beta_{20}, \beta_{21})|\mathbf{y}^R) = \sum_{k=1}^{N}\sum_{i=1}^{n}\left\{p_{ik}(2)^{(c)}(\beta_{20}+\beta_{21}v_{ik})-\log(1+e^{\beta_{20}+\beta_{21}v_{ik}})\right\}.$$

All of the calculations in this section are easily modified for the case that $\mathbf{V} = \sigma^2\mathbf{I}$: simply replace the $\sigma_j^2$'s with $\sigma^2$. For instance, step 2 becomes

$$\sigma^{2(c+1)} = \frac{\sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{ik}(j)^{(c)}\left[y_{ik} - f_j(x_i)^{(c+1)}\right]^2}{Nn}.$$

### *3.4. Choice of the smoothing parameters*

In principal, we can always compute the smoothing parameters by "leave-one-replicate-out" cross-validation. However, for many models, this can be computationally intensive. Fortunately, in the models with $\mathbf{V} = \sigma^2\mathbf{I}$ or $\mathbf{V}_{\mathbf{z}_k} = \mathrm{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$, we can shorten calculations by using Theorem 1 below. In this section, we describe our iterative cross-validation procedure, implemented in our data analysis in Section 6.1. We describe our iterative procedure for $\mathbf{V}_{\mathbf{z}_k} = \mathrm{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$. The steps for $\mathbf{V} = \sigma^2\mathbf{I}$ are the same except with $\hat{\sigma}^2$ replacing the $\hat{\sigma}_j^2$'s .

In our data analysis we set the initial values, the $\lambda_j^{(0)}$'s, to a value that worked well when tested on the data set. We update the $\lambda_j$'s as follows.

1. At iteration $i$, with $\lambda_j = \lambda_j^{(i)}$, $j = 1, \ldots, J$, use the ECM algorithm of Section 3.3 to find the $\hat{p}_{ik}(j)$'s, the $\hat{\sigma}_j^2$'s and the $\hat{f}_j$'s.
2. Discard the $\hat{f}_j$'s from Step 1.
3. Let $\widehat{\mathbf{W}}_{kj}$ be $\mathbf{W}_{kj}$ as defined in (18) but with the $\hat{\sigma}_j^2$'s and $\hat{p}_{ik}(j)$'s replacing the $\sigma_j^2$'s and $p_{ik}(j)$'s. Treat the $\hat{\sigma}_j^2$'s and the $\hat{p}_{ik}(j)$'s and thus the $\widehat{\mathbf{W}}_{kj}$'s as fixed.
4. For $j = 1, \ldots, J$, over a grid of possible $\lambda$ values, set $\lambda_j^{(i+1)}$ as the value of $\lambda$ that minimizes the following leave-one-replicate-out cross-validation criterion:

$$CV_j(\lambda) = \sum_{k=1}^{N}\left[\mathbf{y}_k - \hat{f}_{j\,\lambda}^{(-k)}(\mathbf{x})\right]^T\widehat{\mathbf{W}}_{kj}\left[\mathbf{y}_k - \hat{f}_{j\,\lambda}^{(-k)}(\mathbf{x})\right] \qquad (21)$$

where $\hat{f}_{j\,\lambda}^{(-k)}$ is the function that maximizes

$$S_j^{(-k)}(f_j) = -\frac{1}{2}\sum_{r=1:r\neq k}^{N}\big[\mathbf{y}_r - f_j(\mathbf{x})\big]^T\widehat{\mathbf{W}}_{rj}\big[\mathbf{y}_r - f_j(\mathbf{x})\big] + P(f_j,\lambda) \quad (22)$$

where, for our penalized log-likelihood approach, $P(f_j,\lambda) = -\lambda\int(f_j'')^2 = -\lambda\phi_j^T\mathbf{R}\phi_j$.

5. Repeat steps 1-4 with $\lambda_j = \lambda_j^{(i+1)}$, $j = 1,\ldots,J$, till convergence.

We use the final values of the $\lambda_j$'s to obtain all of the parameter estimates from the ECM algorithm as in Section 3.3.

Finding $\lambda$ that minimizes (21) is computationally intensive. Fortunately, we can prove the following theorem.

**Theorem 1.** *Let $\hat{f}_{1\lambda}, \ldots, \hat{f}_{J\lambda}$ be the maximizers of (15) + (16), with $\mathbf{W}_{kj}$ replaced by $\widehat{\mathbf{W}}_{kj}$. Let $\widehat{\mathbf{H}}_{kj}$ be as in (20), but with $\mathbf{W}_{kj}$ replaced by $\widehat{\mathbf{W}}_{kj}$. Suppose that $\mathbf{I} - \widehat{\mathbf{H}}_{kj}$ is invertible and $\widehat{\mathbf{W}}_{kj}$ is positive definite, $j = 1,\ldots,J$. Then*

$$CV_j(\lambda) = \sum_{k=1}^{N}\Big[(\mathbf{I}-\widehat{\mathbf{H}}_{kj}(\lambda))^{-1}(\hat{f}_{j\,\lambda}(\mathbf{x})-\mathbf{y}_k)\Big]^T\widehat{\mathbf{W}}_{kj}\Big[(\mathbf{I}-\widehat{\mathbf{H}}_{kj}(\lambda))^{-1}(\hat{f}_{j\,\lambda}(\mathbf{x})-\mathbf{y}_k)\Big].$$

The proof follows directly from Lemma 2 in the Appendix B, which holds in a slightly more general setting.

## 4. Standard errors for the parameter estimators of the state process

As in the single realization case presented by De Souza and Heckman (2014) we use the results of Louis (1982) to obtain standard errors for the estimates of the parameters of the state process. For *iid* $z_{ik}$'s we consider $J \geq 2$ possible state values. For Markov $z_{ik}$'s we restrict the possible number of states to $J = 2$ for ease of explanation and computing.

Louis (1982) derived a procedure to obtain the observed information matrix when the maximum likelihood estimates are obtained using the EM algorithm. The procedure requires the computation of the gradient and of the second derivative matrix of the log-likelihood based on the complete data and can be implemented easily within the EM steps. We use Louis's technique with known $\gamma$ to derive the formula for the information matrix associated with $\hat{\alpha}$. We then plug our final estimates, $\hat{\alpha}$ and $\hat{\gamma}$, into this formula, yielding our estimated information matrix:

$$\hat{I} = E_{\alpha,\gamma}(-\mathcal{L}_2''(\alpha)|\mathbf{y}^R) - E_{\alpha,\gamma}(\mathcal{L}_2'(\alpha)\mathcal{L}_2'(\alpha)^T|\mathbf{y}^R)\Big|_{\alpha=\hat{\alpha},\gamma=\hat{\gamma}} \quad (23)$$

where $\mathcal{L}_2(\alpha) = \log p(\mathbf{z}^R|\alpha)$, $\mathcal{L}_2'(\alpha)$ is the gradient vector of $\mathcal{L}_2$ and $\mathcal{L}_2''(\alpha)$ is the associated second derivative matrix. Then $\hat{I}^{-1}$ is the estimated covariance of $\hat{\alpha}$

and the square root of the $jj$th element of $\hat{I}^{-1}$ is the standard error of the $j$th component of $\hat{\alpha}$.

In the next sections we show how to calculate $\hat{I}$ for *iid* and Markov $z_{ik}$'s. Using similar techniques we also obtained standard errors for the intercept and slope parameters for the case when $J = 2$ and the $z_{ik}$'s are independent with distribution depending on only one covariate.

### *4.1. Standard errors: iid $z_{ik}$'s*

Setting the parameters in $\alpha$ equal to $p_1, \ldots, p_{J-1}$ and noting that $p_J = 1 - \sum_{m=1}^{J-1} p_m$, consider the $(J-1)\times(J-1)$ matrix $\mathrm{E}_\alpha(-\mathcal{L}_2''(\alpha)|\mathbf{y}^R)$ in (23) evaluated at $\alpha = \hat{\alpha}$ and $\gamma = \hat{\gamma}$. One can show its $jl$th entry, for $j \neq l$, is equal to $Nn/\hat{p}_J$, where $\hat{p}_J = 1 - \sum_{m=1}^{J-1} \hat{p}_m$. Its $jj$th entry is

$$Nn \times \left( \frac{1}{\hat{p}_j} + \frac{1}{\hat{p}_J} \right),$$

Recall that $n_{\mathbf{s},j} = \sum_{i=1}^{n} \mathrm{I}\{s_i = j\}$. One can also show that the $(J-1)\times(J-1)$ matrix $\mathrm{E}_\alpha(\mathcal{L}_2'(\alpha)\mathcal{L}_2'(\alpha)^T|\mathbf{y}^R)$ in (23) evaluated at $\alpha = \hat{\alpha}$ and $\gamma = \hat{\gamma}$ has off diagonal elements $jl$ equal to

$$\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \left( \frac{n_{\mathbf{s},j}}{\hat{p}_j} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right) \left( \frac{n_{\mathbf{s},l}}{\hat{p}_l} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right)$$

$$- \sum_{k=1}^{N} \left[ \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \left( \frac{n_{\mathbf{s},j}}{\hat{p}_j} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right) \times \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \left( \frac{n_{\mathbf{s},l}}{\hat{p}_l} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right) \right]$$

and $j$th diagonal element given by

$$\sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \left( \frac{n_{\mathbf{s},j}}{\hat{p}_j} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right)^2 - \sum_{k=1}^{N} \left[ \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \left( \frac{n_{\mathbf{s},j}}{\hat{p}_j} - \frac{n_{\mathbf{s},J}}{\hat{p}_J} \right) \right]^2,$$

where $\hat{p}_k(\mathbf{s}) = p(\mathbf{z}_k = s|\mathbf{y}_k, \hat{\alpha}, \hat{\gamma})$.

### *4.2. Standard errors: Markov $z_{ik}$'s*

For Markov $z_{ik}$'s we show how to obtain standard errors for the estimates of the parameters of the $z$ process for $J = 2$ possible state values. We apply Louis's method to obtain standard errors for $\hat{\pi}_1$, $\hat{a}_{12}$ and $\hat{a}_{21}$ simultaneously.

Recall that $n_{\mathbf{s},lj} = \sum_{i=2}^{n} \mathrm{I}(s_{i-1} = l, s_i = j)$. The $3\times3$ matrix $\mathrm{E}_\alpha(-\mathcal{L}''(\alpha)|\mathbf{y}^R)$ in (23) evaluated at $\alpha = \hat{\alpha}$ and $\gamma = \hat{\gamma}$ is diagonal with entries

$$\frac{N}{\hat{\pi}_1(1-\hat{\pi}_1)}, \quad \frac{1}{\hat{a}_{12}(1-\hat{a}_{12})} \sum_{k=1}^{N} \sum_{\text{all } \mathbf{s}\text{'s}} \hat{p}_k(\mathbf{s}) \sum_{i=2}^{n} \mathrm{I}(s_{i-1} = 1) \quad \text{and}$$

$$\frac{1}{\hat{a}_{21}(1-\hat{a}_{21})}\sum_{k=1}^{N}\sum_{\text{all }\mathbf{s}\text{'s}}\hat{p}_k(\mathbf{s})\sum_{i=2}^{n}\mathrm{I}(s_{i-1}=2).$$

The calculation of the $3\times3$ symmetric matrix $\mathrm{E}_\alpha(\mathcal{L}_2'(\alpha)\mathcal{L}_2'(\alpha)^T|\mathbf{y}^R)$ evaluated at $\alpha=\hat{\alpha}$ and $\gamma=\hat{\gamma}$ is straightforward but long. The matrix entries are given in Section 2 of the Supplementary Material.

## 5. Simulation studies

We carry out three simulation studies and consider that the $z_{ik}$'s can take values 1 or 2. For each simulation study 300 independent data sets are generated, each with $N=100$ replicates. In study 1, $z_{1k},\ldots,z_{nk}$ are *iid* and $\mathbf{V}$ follows the homogeneous random intercept model as in (3). In study 2, $z_{1k},\ldots,z_{nk}$ follow a Markov structure and $\mathbf{V}$ also follows the homogeneous random intercept model. In study 3, $z_{1k},\ldots,z_{nk}$ are independent, but with the distribution of $z_{ik}$ depending on a univariate covariate, $v_{ik}$. In this third study, we take $\mathbf{V}=\sigma^2\mathbf{I}$. In all three studies we use the same vector of evaluation points $\mathbf{x}$ and the same true functions $f_1$ and $f_2$. The vector $\mathbf{x}=(x_1,\ldots,x_n)^T$ consists of $n=10$ equally spaced points, $1,12,23,\ldots,89,100$. The true function $f_2$ is the same as used by De Souza and Heckman (2014) in their simulation studies. The true function $f_1$ is simply $f_2-0.1$. In the third study, in each simulated data set, we generate $v_{ik}, k=1,\ldots,n, i=1,\ldots,N$. Figure 3 presents examples of a data set from each simulation study with 20 of their 100 replicates.

For Simulations 1 and 2 we generate each simulated data set as follows.

1. Generate the $z_{ik}$'s according to the specified model - *iid* for Simulation 1, Markov for Simulation 2. For the *iid* model, we set $p_1=p(z_{ik}=1)=0.5$. For Markov $z_{ik}$'s, we set transition probabilities $a_{12}=p(z_i=2|z_{i-1}=1)=0.3$ and $a_{21}=p(z_i=1|z_{i-1}=2)=0.4$ and initial probabilities $\pi_1=\pi_2=0.5$.
2. Generate the $y_{ik}$'s according to the homogeneous random intercept model of Section 2.1 with $\tau^2=10^{-4}$ and $\sigma^2=10^{-5}$.
3. Repeat steps 1 and 2 $N=100$ times obtaining a data set of 100 replicates.

For Simulation 3 we generate each simulated data as follows.

1. Generate $v_{ik}$'s *iid* $N(0,1)$.
2. Generate the $z_{ik}$'s such that $\mathrm{P}\{z_{ik}=1|v_{ik}\}=1/[1+\exp(\beta_0+\beta_1 v_{ik})]$ and so $\log[p_2(v_{ik})/p_1(v_{ik})]=\beta_0+\beta_1 v_{ik}$. We set $\beta_0=2$ and $\beta_1=5$.
3. Generate the $y_{ik}$'s as follows. If $z_{ik}=1$ then $y_{ik}=f_1(x_i)+e_{ik}$. If $z_{ik}=2$ then $y_{ik}=f_2(x_i)+e_{ik}$. The $\epsilon_{ik}$'s are *iid* $N(0,\sigma^2)$. We set $\sigma^2=10^{-5}$.
4. Repeat steps 1, 2 and 3 $N=100$ times obtaining a data set of 100 replicates.

We analyze the data using the proposed EM algorithm. We set initial parameter values to be the true parameter values to speed up computation. We

did try initial values that were different than the true parameter values and the EM algorithm also converged, but it took longer than when starting from the truth, as expected.

The values of $\lambda_1$ and $\lambda_2$ are fixed and equal to $10^{-4}$ in all simulation studies. The value $10^{-4}$ was chosen by eye when tested on few simulated data sets.

### 5.1. Results

The three plots in Figure 3 show the fitted values $\hat{f}_1(\mathbf{x})$ and $\hat{f}_2(\mathbf{x})$ (dashed curves) for a simulated data set from each of simulation studies 1, 2 and 3.

We assess the quality of the estimated functions via the pointwise empirical mean squared error (EMSE) as shown in Figure 4. We observe that in all simulation studies $\hat{f}_1(\mathbf{x})$ and $\hat{f}_2(\mathbf{x})$ produce very small values of EMSE ($< 2 \times 10^{-5}$). However, Simulation 3's EMSE values for $\hat{f}_1(\mathbf{x})$ are larger than in Simulations 1 and 2.

Table 1 presents the mean and standard deviation of all 300 estimates of $\sigma^2$ and $\tau^2$ for Simulations 1 and 2. Table 2 presents the mean and standard deviation of all 300 estimates of $\sigma^2$ for Simulation 3. We can observe that in all three simulation studies we are slightly underestimating the values of these variance parameters. This may be due to the challenges of correctly adjusting the degrees of freedom in the estimates, in order to account for the estimation of the $f_j$'s.

Table 3 contains the mean and the standard deviation of the estimates of the parameters of the latent process for each simulation study, along with the averages of our proposed standard errors. Note that the standard deviations of the estimates are close to the values of the means of the proposed standard errors (s.e.'s), as desired. Table 3 also shows the empirical coverage percentages of both a 90% and a 95% confidence interval. We consider confidence intervals of the form "mean of the parameter estimates $\pm z_{\alpha/2} \times$ proposed s.e.", where $z_{\alpha/2}$ is the $\alpha/2$ quantile of a standard normal distribution with $\alpha = 0.1$ and 0.05. The empirical coverage percentages for all three simulation studies are very close to the true level of the corresponding confidence interval.

## 6. Analysis of the power usage data

The data shown in Figure 2 consist of daytime hourly power usage of a building from 9am to 4pm ($n = 8$ observations in a day) on $N = 44$ business days in June and July 2009. For the same days and hours we also have available the temperature near the building. We apply our proposed methodology to these data treating each day as a replicate and modelling power usage as arising from $J = 2$ functions, one function giving power usage when the chiller is off ($j = 1$), and the other function giving power usage when the chiller is on ($j = 2$). In Section 6.1 we present the results assuming the covariance matrix $\mathbf{V}$ is diagonal and in Section 6.2 we present the results when we assume $\mathbf{V}$ is generated by the non-homogeneous random intercept model as in (4).

### 6.1. Results: diagonal $\mathbf{V}$

In this section we consider two models for $\mathbf{V}$: $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{V} = \mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$. We use the ECM algorithm described in Section 3.3 to estimate the model parameters considering *iid* $z_{ik}$'s, Markov $z_{ik}$'s and $z_{ik}$'s that are independent with distribution depending on temperature. The smoothing parameters, the $\lambda_j$'s, are chosen by cross-validation as described in detail in Section 3.4.

Figures 5a and 5b present the fitted functions for *iid* hidden states $z_{ik}$'s when we assume $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{V}_{\mathbf{z}_k}$, respectively. We can observe that the fitted curves are very similar in the two figures. The estimated curve giving power usage when the chiller is on, obtained assuming $\mathbf{V} = \sigma^2\mathbf{I}$, is slightly smoother than the one obtained assuming $\mathbf{V}_{\mathbf{z}_k}$. Table 4 presents the parameter estimates and chosen $\lambda_j$'s. We can see that the estimates of $p_j = p(z_{ik} = j)$ from the two models for $\mathbf{V}$ agree within the reported standard errors. We also observe in the lower half of the Table that the estimated variance when the chiller is on is much higher than when the chiller is off.

Figures 6a and 6b present the fitted curves for Markov $z_{ik}$'s when we assume $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{V}_{\mathbf{z}_k}$, respectively. As in the *iid* case, the fitted curve giving power usage when the chiller is on obtained assuming $\mathbf{V} = \sigma^2\mathbf{I}$ is slightly smoother than the one obtained assuming $\mathbf{V}_{\mathbf{z}_k}$. Table 5 provides information on the estimated model parameters and the chosen smoothing parameters. As in the *iid* case, the estimated variance when the chiller is on is much higher than when the chiller is off. We observe that the estimates of $a_{21}$, the transition probability from "chiller on" to "chiller off", are very small or equal to zero. Any estimate of $a_{21}$ is expected to be small, as there is only one replicate in the data set where we see this transition. The estimate of zero is reasonable when we assume different variances; $\hat{a}_{21}$ is zero because the transition happens gradually, which our model does not allow, and the method incorrectly classifies all observations as coming from the condition "chiller on", failing to detect the transition. This replicate is the green curve in Figure 6b.

Figure 7 presents the fitted curves when we assume the $z_{ik}$'s are independent with distribution depending on temperature via the following logistic regression model:

$$\log \frac{p(\text{chiller on} \mid \text{temperature})}{p(\text{chiller off} \mid \text{temperature})} = \beta_0 + \beta_1 \text{ temperature}.$$

Table 6 shows the corresponding estimated model parameters assuming $\mathbf{V}_{\mathbf{z}_k}$ along with the chosen smoothing parameters, the $\lambda_j$'s. We observe in Table 6 that the standard error for $\hat{\beta}_1$ is very small and by considering a confidence interval of the form $\hat{\beta}_1 \pm 1.96 \times \text{s.e.}(\hat{\beta}_1)$ we conclude that the coefficient $\beta_1$ is statistically significant.

### 6.2. Results: correlated observations generated by the non-homogeneous random intercept model

In the analyses of Section 6.1, we see that the variability in energy consumption when the chiller is on is higher than when the chiller is off. Thus, models such as $\mathbf{V} = \sigma^2\mathbf{I}$ or $\mathbf{V}$ following the homogeneous random intercept model may not be appropriate. Therefore, to model this heterogeneity in variance and the correlation between observations, we fit the proposed switching nonparametric regression model to the power usage data assuming the covariance matrix $\mathbf{V}$ is generated by the non-homogeneous random intercept model as in (4). We use the ECM algorithm described in Section 3 and in Section 1 of the Supplementary Material to obtain the parameter estimates. We conduct the analysis assuming the hidden states $z_{ik}$'s are *iid*. The smoothing parameters, the $\lambda_j$'s, are fixed and chosen by eye.

Table 7 presents the parameter estimates. We observe that the estimates of $p_1$ and $p_2$ in Table 7 agree within the reported standard errors with the estimates obtained in Table 4 where we assume the observations are uncorrelated. Figure 8 shows the corresponding fitted curves. We can observe that the fitted function corresponding to the condition "chiller on" is lower than that in Figures 5 to 7. The non-homogeneous random intercept model appears to "explain" days of high power usage by a larger variability of the "chiller on" random intercept. Thus the replicates with very high power usage have less of an impact on the final fitted "chiller on" curve.

## 7. Discussion

We have introduced a method for the analysis of data arising from random samples of a process with a complex structure. The structure depends on a latent state process where each state corresponds to a true smooth regression function. The estimation techniques and standard error calculations were developed for several specific cases of state processes and error covariances. While the models we considered may not capture all of the dependencies in a data set, our techniques and ideas should carry over to more complex latent state processes and richer time series modelling of the error process. Further useful extensions might incorporate a dependence among replicates; for instance, in studying energy consumption of several buildings, one would want to incorporate a random "building" effect.

## Appendix A: Forms of $p_{ik}(j)^{(c)}$ in the EM algorithm when V is diagonal

When the $z_{ik}$'s are independent, similarly to the one realization case of De Souza and Heckman (2014), we obtain

$$\begin{aligned} p_{ik}(j)^{(c)} &= p(z_{ik}=j|y_{ik},\theta^{(c)}) \\ &= \frac{p(y_{ik}|z_{ik}=j,f_j(x_i)^{(c)},\sigma_j^{2\,(c)})\times\tilde{p}_j^{(c)}}{\sum_{l=1}^{J}p(y_{ik}|z_{ik}=l,f_l(x_i)^{(c)},\sigma_j^{2\,(c)})\times\tilde{p}_l^{(c)}} \end{aligned}$$

with $p(y_{ik}|z_{ik}=j,f_j(x_i)^{(c)},\sigma_j^{2\,(c)})$ easily calculated with our normality assumption. Here, $\tilde{p}_l^{(c)}=p_l^{(c)}$ when the $z_{ik}$'s are *iid*, and when $z_{ik}$ depends on the covariate $\mathbf{v}_{ik}$, $\tilde{p}_l^{(c)}=p(z_{ik}=l|\mathbf{v}_{ik})$.

When the $z_{ik}$'s are Markov,

$$p_{ik}(j)^{(c)}=\frac{\delta_{ik}(j)^{(c)}\varphi_{ik}(j)^{(c)}}{\sum_{l=1}^{J}\delta_{ik}(l)^{(c)}\varphi_{ik}(l)^{(c)}},$$

where $\delta_{ik}(j)^{(c)}=p(y_{1k},\ldots,y_{ik},z_{ik}=j|\theta^{(c)})$ and $\varphi_{ik}(j)^{(c)}=p(y_{(i+1)k},\ldots,y_{nk}|z_{ik}=j,\theta^{(c)})$ are obtained using, respectively, the forward and backward procedures proposed by Baum et al. (1970).

## Appendix B: Proof of Theorem 1

Theorem 1 is based on the following lemmas, which frame the problem for fixed $j$ and fixed $\lambda$ (so these are dropped in notation) and with general matrices $\mathcal{W}_r$, $r=1,\ldots,N$. Lemma 1 holds for general penalties, while Lemma 2 places further restrictions, restrictions that hold in our setting. Throughout, we assume that all maximizers exist.

Let $\hat{f}^{(-k)}$ maximize

$$S^{(-k)}(f)=-\frac{1}{2}\sum_{r=1;r\neq k}^{N}\big[\mathbf{y}_r-f(\mathbf{x})\big]^T\mathcal{W}_r\big[\mathbf{y}_r-f(\mathbf{x})\big]+P(f).$$

**Lemma 1.** *Let* $\hat{f}^{(*k)}$ *maximize*

$$\begin{aligned} S^{(*k)}(f) &= -\frac{1}{2}[\hat{f}^{(-k)}(\mathbf{x})-f(\mathbf{x})]^T\mathcal{W}_k[\hat{f}^{(-k)}(\mathbf{x})-f(\mathbf{x})] \\ &\quad -\frac{1}{2}\sum_{r=1,r\neq k}^{N}[\mathbf{y}_r-f(\mathbf{x})]^T\mathcal{W}_r[\mathbf{y}_r-f(\mathbf{x})] \quad + \quad P(f). \end{aligned}$$

*If* $\mathcal{W}_k$ *is positive definite then* $\hat{f}^{(-k)}(\mathbf{x})=\hat{f}^{(*k)}(\mathbf{x})$.

*Proof of Lemma 1.* For simplicity let $k = 1$. We want to show that $\hat{f}^{(-1)} = \hat{f}^{(*1)}$. We know $\hat{f}^{(-1)}$ maximizes $S^{(-1)}(f)$ and, therefore,

$$S^{(-1)}(\hat{f}^{(-1)}) - S^{(-1)}(\hat{f}^{(*1)}) \geq 0.$$

We also know that $\hat{f}^{(*1)}$ maximizes $S^{(*1)}(f)$. Thus, $S^{(*1)}(\hat{f}^{(*1)}) - S^{(*1)}(\hat{f}^{(-1)}) \geq 0$, that is,

$$\begin{aligned} &-\frac{1}{2}\big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big]^T \mathcal{W}_1 \big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big] \\ &-\frac{1}{2}\sum_{r=2}^{N} \big[\mathbf{y}_r - \hat{f}^{(*1)}(\mathbf{x})\big]^T \mathcal{W}_r \big[\mathbf{y}_r - \hat{f}^{(*1)}(\mathbf{x})\big] \;+\; P(\hat{f}^{(*1)}) \\ &+\frac{1}{2}\sum_{r=2}^{N} \big[\mathbf{y}_r - \hat{f}^{(-1)}(\mathbf{x})\big]^T \mathcal{W}_r \big[\mathbf{y}_r - \hat{f}^{(-1)}(\mathbf{x})\big] \;-\; P(\hat{f}^{(-1)}) \geq 0, \end{aligned}$$

that is,

$$-\frac{1}{2}\big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big]^T \mathcal{W}_1 \big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big] \geq S^{(-1)}(\hat{f}^{(-1)}) - S^{(-1)}(\hat{f}^{(*1)}) \geq 0,$$

which implies that

$$\big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big]^T \mathcal{W}_1 \big[\hat{f}^{(-1)}(\mathbf{x}) - \hat{f}^{(*1)}(\mathbf{x})\big] \leq 0,$$

and, because $\mathcal{W}_1$ is positive definite, $\hat{f}^{(-1)}(\mathbf{x}) = \hat{f}^{(*1)}(\mathbf{x})$. □

**Lemma 2.** *Suppose that $\mathcal{W}_k$ is positive definite, $k = 1, \ldots, N$. Let $\hat{f}$ maximize*

$$S(f) = -\frac{1}{2}\sum_{k=1}^{N} \big[\mathbf{y}_k - f(\mathbf{x})\big]^T \mathcal{W}_k \big[\mathbf{y}_k - f(\mathbf{x})\big] \;+\; P(f).$$

*If there exist matrices $\mathcal{H}_k$, $k = 1, \ldots, N$, not depending on the $\mathbf{y}_r$'s, such that $\hat{f}(\mathbf{x}) = \sum_{k=1}^{N} \mathcal{H}_k \mathbf{y}_k$, then*

$$(\mathbf{I} - \mathcal{H}_k)\,[\hat{f}^{(-k)}(\mathbf{x}) - \mathbf{y}_k] = \hat{f}(\mathbf{x}) - \mathbf{y}_k.$$

*Proof of Lemma 2.* Note that $\hat{f}^{(*k)}$, as defined in Lemma 1, is the maximizer of $S$ with $\mathbf{y}_k$ replaced by $\hat{f}^{(-k)}$. By the assumption of the form of the maximizer of $S$, $\hat{f}^{(*k)}(\mathbf{x})$ can be written as

$$\begin{aligned} \hat{f}^{(*k)}(\mathbf{x}) &= \sum_{r=1:r\neq k}^{N} \mathcal{H}_r \mathbf{y}_r + \mathcal{H}_k \hat{f}^{(-k)}(\mathbf{x}) \\ &= \hat{f}(\mathbf{x}) - \mathcal{H}_k \mathbf{y}_k + \mathcal{H}_k \hat{f}^{(-k)}(\mathbf{x}). \end{aligned}$$

From Lemma 1 we know $\hat{f}^{(-k)}(\mathbf{x}) = \hat{f}^{(*k)}(\mathbf{x})$. Thus,

$$\hat{f}^{(-k)}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \mathcal{H}_k \mathbf{y}_k + \mathcal{H}_k \hat{f}^{(-k)}(\mathbf{x}).$$

Now subtracting $\mathbf{y}_k$ from both sides of this equation, we obtain

$$(\mathbf{I} - \mathcal{H}_k)[\hat{f}^{(-k)}(\mathbf{x}) - \mathbf{y}_k] = \hat{f}(\mathbf{x}) - \mathbf{y}_k.$$

□

## References


Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41** 164–171.

Brown, M., Barrington-Leigh, C. and Brown, Z. (2012). Kernel regression for real-time building energy analysis. *Journal of Building Performance Simulation* **5** 263–276.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Verlag.

De Souza, C. P. E. and Heckman, N. E. (2014). Switching nonparametric regression models. *Journal of Nonparametric Statistics* **26** 617–637.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39** 1–38.

Langrock, R., Kneib, T., Glennie, R. and Michelot, T. (2015). Markov-switching generalized additive models. *Statistics and Computing* 1-12.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B* **44** 226–233.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. 2nd Ed., Wiley New York.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80** 267-278.

Ou, X. and Martin, E. (2008). Batch process modelling with mixtures of Gaussian processes. *Neural Computing & Applications* **17** 471–479.

Rasmussen, C. E. and Ghahramani, Z. (2002). Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference* **2** 881–888. The MIT Press.

Tresp, V. (2001). Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference* 654–660. The MIT Press.

TABLE 1
*Simulations 1 and 2: estimates of the covariance parameters $\sigma^2$ (true value $= 1 \times 10^{-5}$) and $\tau^2$ (true value $= 1 \times 10^{-4}$).*

| Simulation | mean$\times 10^5$ (SD* $\times 10^5$) | mean$\times 10^4$ (SD* $\times 10^4$) |
|---|---|---|
| 1 | 0.978 (0.046) | 0.977 (0.152) |
| 2 | 0.978 (0.045) | 0.977 (0.152) |

*SD = standard deviation.

TABLE 2
*Simulation 3: estimates of $\sigma^2$ (true value $= 5 \times 10^{-5}$).*

| Simulation | mean$\times 10^5$ (SD* $\times 10^5$) |
|---|---|
| 3 | 4.919 (0.238) |

*SD = standard deviation.

TABLE 3
*Simulations 1, 2 and 3: estimates of the parameters of the z process.*

| | | | | empirical coverage | |
|---|---|---|---|---|---|
| Sim'n | $z$ parameters | mean (SD*) | mean of s.e.'s | 90% | 95% |
| 1 | $p_1 = 0.5$ | 0.499 (0.016) | 0.016 | 90.3% | 95.7% |
| 2 | $\pi_1 = 0.5$ | 0.502 (0.050) | 0.050 | 89.7% | 95.7% |
| | $a_{12} = 0.3$ | 0.300 (0.021) | 0.020 | 90.0% | 94.3% |
| | $a_{21} = 0.4$ | 0.401 (0.024) | 0.025 | 89.7% | 95.3% |
| 3 | $\beta_0 = 2$ | 2.010 (0.173) | 0.176 | 91.0% | 96.7% |
| | $\beta_1 = 5$ | 5.047 (0.357) | 0.364 | 90.6% | 94.3% |

*SD = standard deviation.

TABLE 4
*Data analysis results for iid $z_{ik}$'s for $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{V}_{\mathbf{z}_k} = diag(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$, with corresponding fitted curves in Figures 5a and 5b, respectively.*

| | curve (chiller condition, $j$) | $\hat{\sigma}^2$ | $\hat{p}_j$ (s.e.) | $\lambda_j$ |
|---|---|---|---|---|
| $\mathbf{V} = \sigma^2\mathbf{I}$ | black (off, $j = 1$) | 103.5 | 0.665 (0.025) | 0.020 |
| | red (on, $j = 2$) | | 0.335 (0.025) | 0.078 |
| $\mathbf{V}_{\mathbf{z}_k} =$ | black (off, $j = 1$) | 12.7 | 0.658 (0.025) | 0.073 |
| $diag(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$ | red (on, $j = 2$) | 355.4 | 0.342 (0.025) | 0.006 |

TABLE 5
*Data analysis results for Markov $z_{ik}$'s, for $\mathbf{V} = \sigma^2\mathbf{I}$ and $\mathbf{V}_{\mathbf{z}_k} = diag(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$, with corresponding fitted curves in Figures 6a and 6b.*

| | curve (chiller condition, $j$) | $\hat{\sigma}^2$ | $\hat{\pi}_j$ (s.e.) | $\hat{a}_{12}$ (s.e.) | $\hat{a}_{21}$ (s.e.) | $\lambda_j$ |
|---|---|---|---|---|---|---|
| $\mathbf{V} = \sigma^2\mathbf{I}$ | black (off, $j = 1$) | 103.1 | 0.705 (0.069) | 0.024 | 0.00991 | 0.019 |
| | red (on, $j = 2$) | | 0.295 (0.069) | (0.011) | (0.00986) | 0.084 |
| $\mathbf{V}_{\mathbf{z}_k} =$ | black (off, $j = 1$) | 12.2 | 0.682 | 0.015 | $< 10^{-16}$ | 0.050 |
| $diag(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$ | red (on, $j = 2$) | 400.2 | 0.318 | | | 0.006 |

TABLE 6
*Data analysis results for $z_{ik}$'s with distribution depending on a covariate (temperature) and $\mathbf{V}_{\mathbf{z}_k} = diag(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$ with corresponding fitted curves in Figure 7.*

| curve (chiller condition, $j$) | $\hat{\sigma}^2_j$ | $\hat{\beta}$ (s.e.) | $\lambda_j$ |
|---|---|---|---|
| black (off, $j = 1$) | 17.9 | $\hat{\beta}_0 = -13.015$ (1.412) | 0.115 |
| red (on, $j = 2$) | 273.9 | $\hat{\beta}_1 = 0.607$ (0.068) | 0.030 |

TABLE 7
*Data analysis results for iid $z_{ik}$'s and $\mathbf{V}$ depending on the hidden states generated by a random intercept model with $\lambda_1 = \lambda_2 = 0.1$ and corresponding curves in Figure 8.*

| curve (chiller condition, $j$) | $\hat{\sigma}^2$ | $\hat{\tau}^2_1$ | $\hat{\tau}^2_2$ | $\hat{p}_j$ (s.e.) |
|---|---|---|---|---|
| black (off, $j = 1$) | 12.4 | 9.3 | 472.0 | 0.672 (0.025) |
| red (on, $j = 2$) | | | | 0.328 (0.025) |

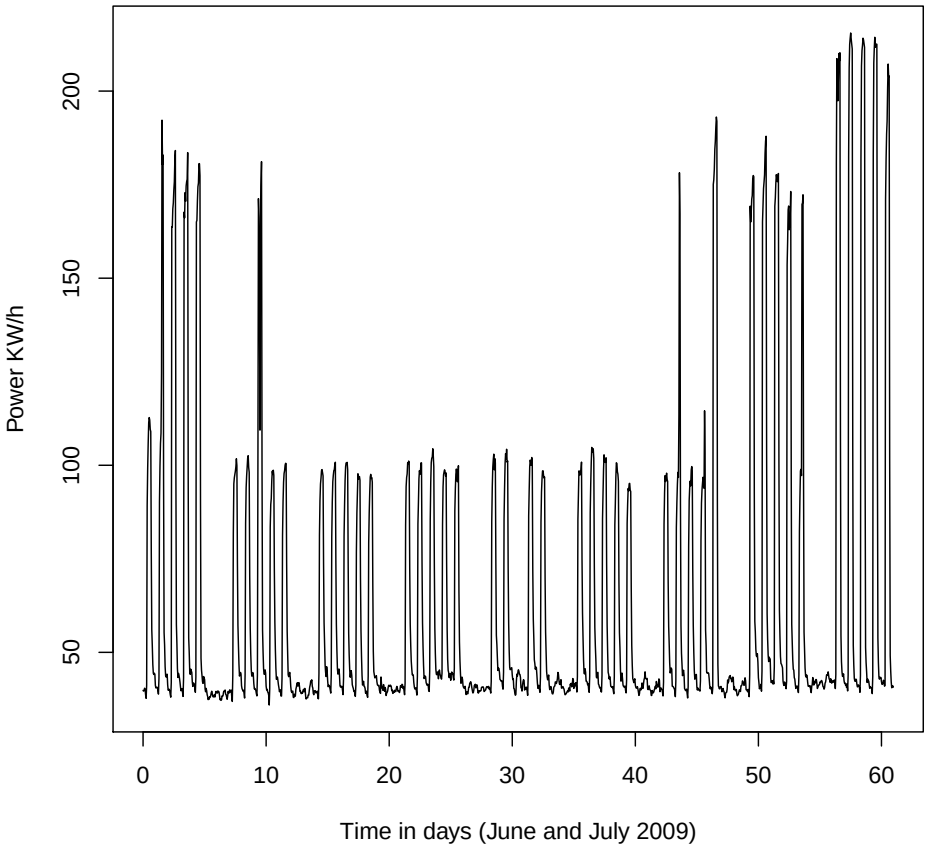



Fig 1: Power usage in June and July 2009 in a building monitored by Pulse Energy.

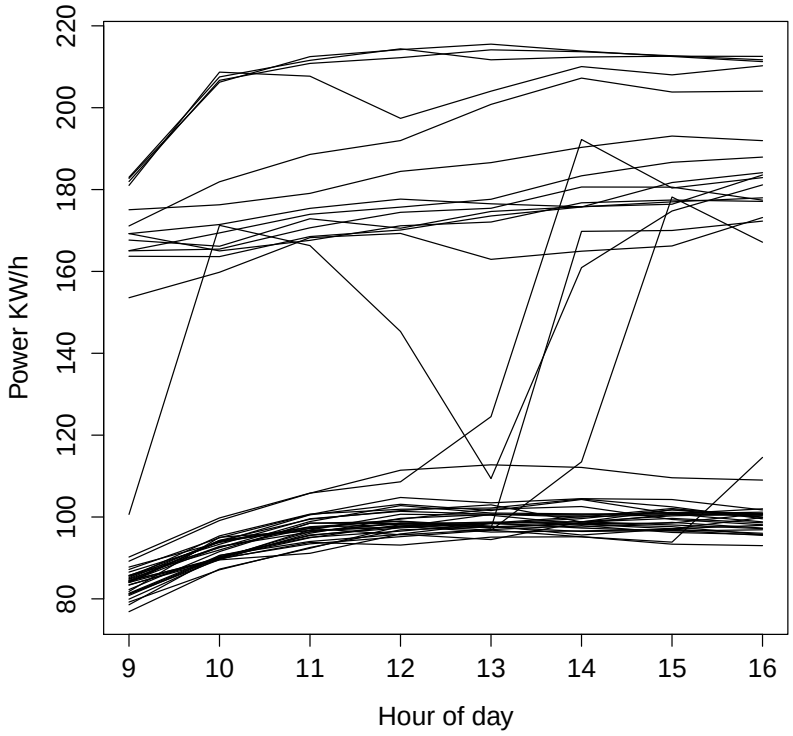
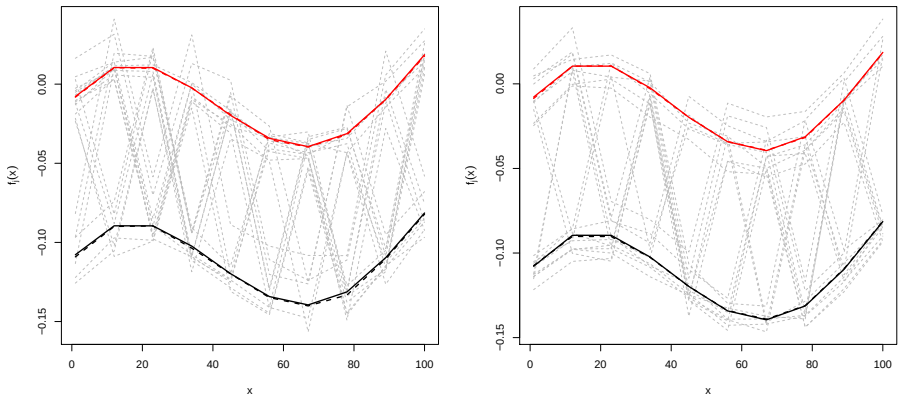
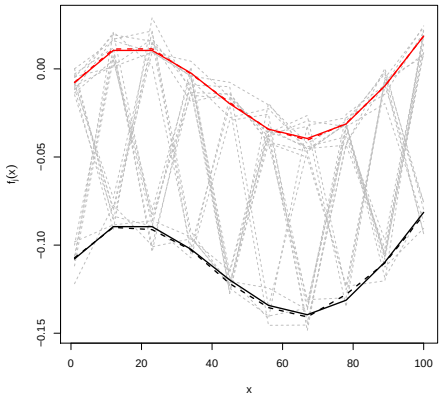



Fig 2: Daytime power usage from 9am to 4pm on business days in June and July 2009 in the same building as Figure 1. Each curve corresponds to a different day.



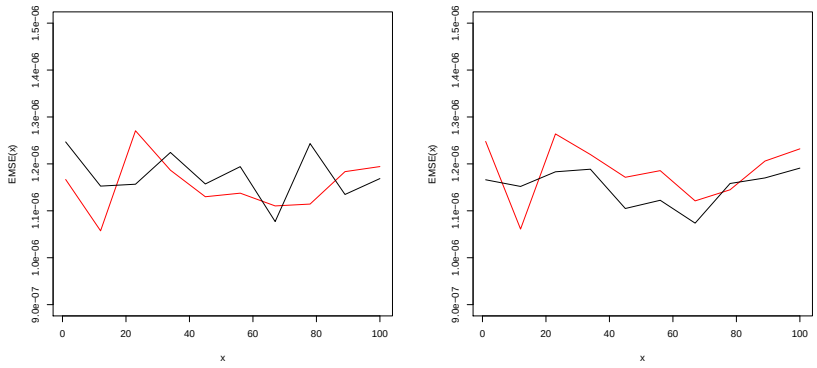

(a) Simulation 1: *iid* $z_{ik}$'s

(b) Simulation 2: Markov $z_{ik}$'s




(c) Simulation 3: covariate dependent $z_{ik}$'s

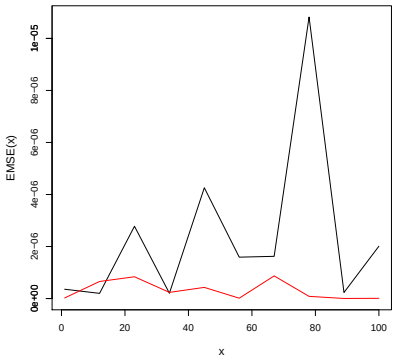
Fig 3: Example of simulated data along with $\hat{f}_1(\mathbf{x})$ and $\hat{f}_2(\mathbf{x})$ for each simulation study. The gray dashed curves correspond to 20 out of the 100 generated replicates. The black and red solid curves correspond to the true functions $f_1$ and $f_2$, respectively, evaluated only at $\mathbf{x}$. The black and red dashed curves correspond to $\hat{f}_1(\mathbf{x})$ and $\hat{f}_2(\mathbf{x})$, respectively.




(a) Simulation 1

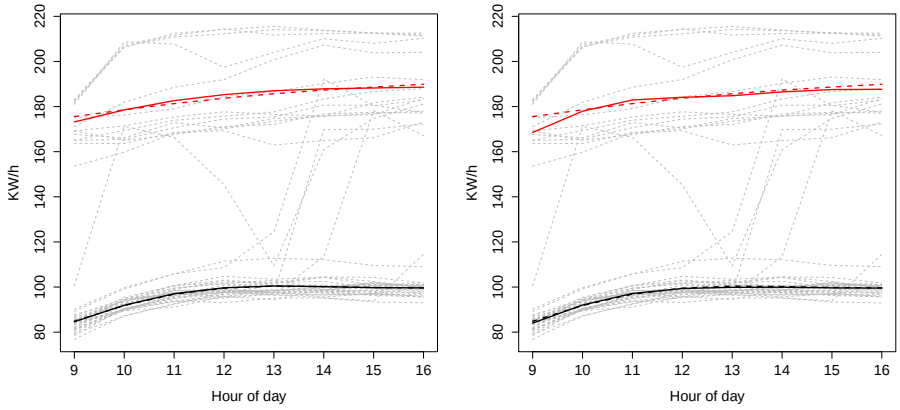



(b) Simulation 2



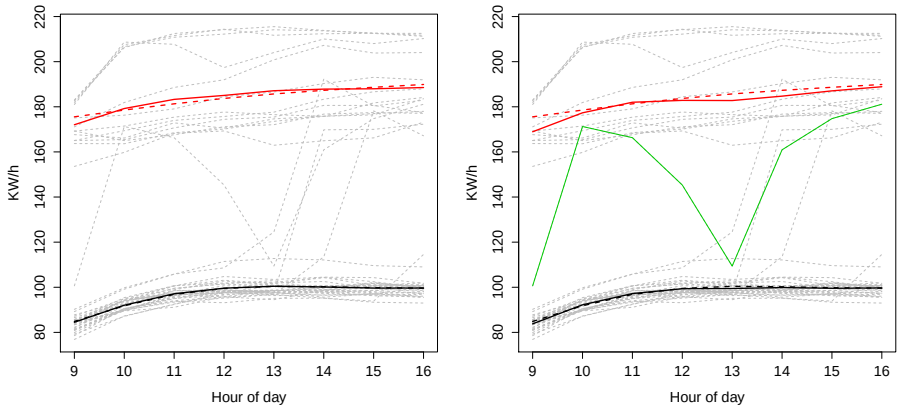

(c) Simulation 3

Fig 4: EMSE of $\hat{f}_1(\mathbf{x})$ (black curves) and $\hat{f}_2(\mathbf{x})$ (red curves) for all three simulation studies.




(a) (b)

Fig 5: Building daytime power usage. Fitted function estimates (solid curves) assuming *iid* $z_{ik}$'s. In (a) we consider $\mathbf{V} = \sigma^2\mathbf{I}$ and in (b) $\mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$. The gray dashed curves correspond to the replicates. The red and black dashed curves are the initial function estimates. The colors red and black correspond to the condition chiller on and off, respectively.




(a) (b)

Fig 6: Building daytime power usage. Fitted function estimates (solid curves) assuming Markov $z_{ik}$'s. In (a) we consider $\mathbf{V} = \sigma^2\mathbf{I}$ and in (b) $\mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$. Components of the plot are as in Figure 5a. The green curve in (b) corresponds to the replicate where there is a transition from chiller on to off.

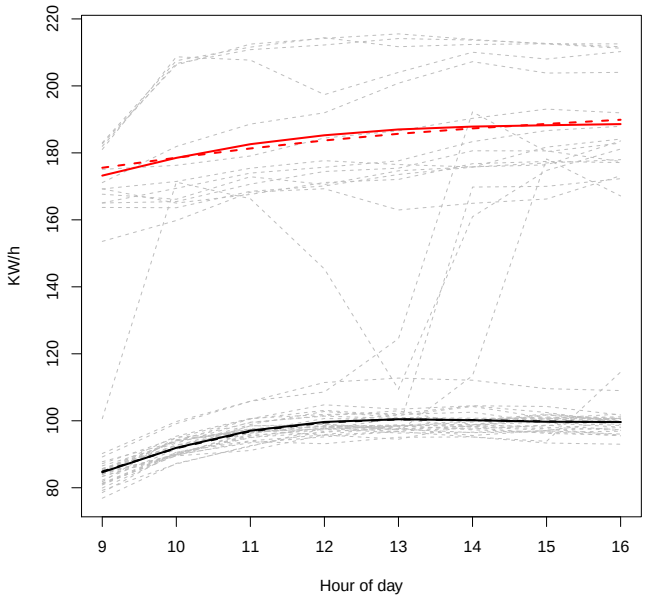



Fig 7: Building daytime power usage. Fitted function estimates assuming the $z_{ik}$'s are independent with distribution depending on temperature. $\mathbf{V}_{\mathbf{z}_k} = \text{diag}(\sigma^2_{z_{1k}}, \ldots, \sigma^2_{z_{nk}})$. Components of the plot are as in Figure 5a.

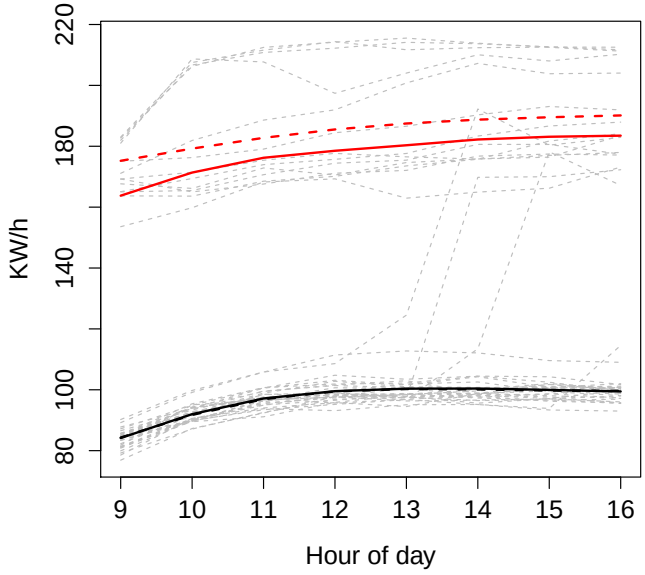



Fig 8: Building daytime power usage. Fitted function estimates assuming *iid* $z_{ik}$'s and $\mathbf{V}$ generated by a non-homogeneous random intercept model. Components of the plot are as in Figure 5a.